Introduction to Protein Structure

Function, evolution & experimental methods

Thomas Nordahl Petersen
CFB Center For Biosustainability,
DTU Lyngby

Thomas Holberg Blicher
NNF Center for Protein Research
University of Copenhagen

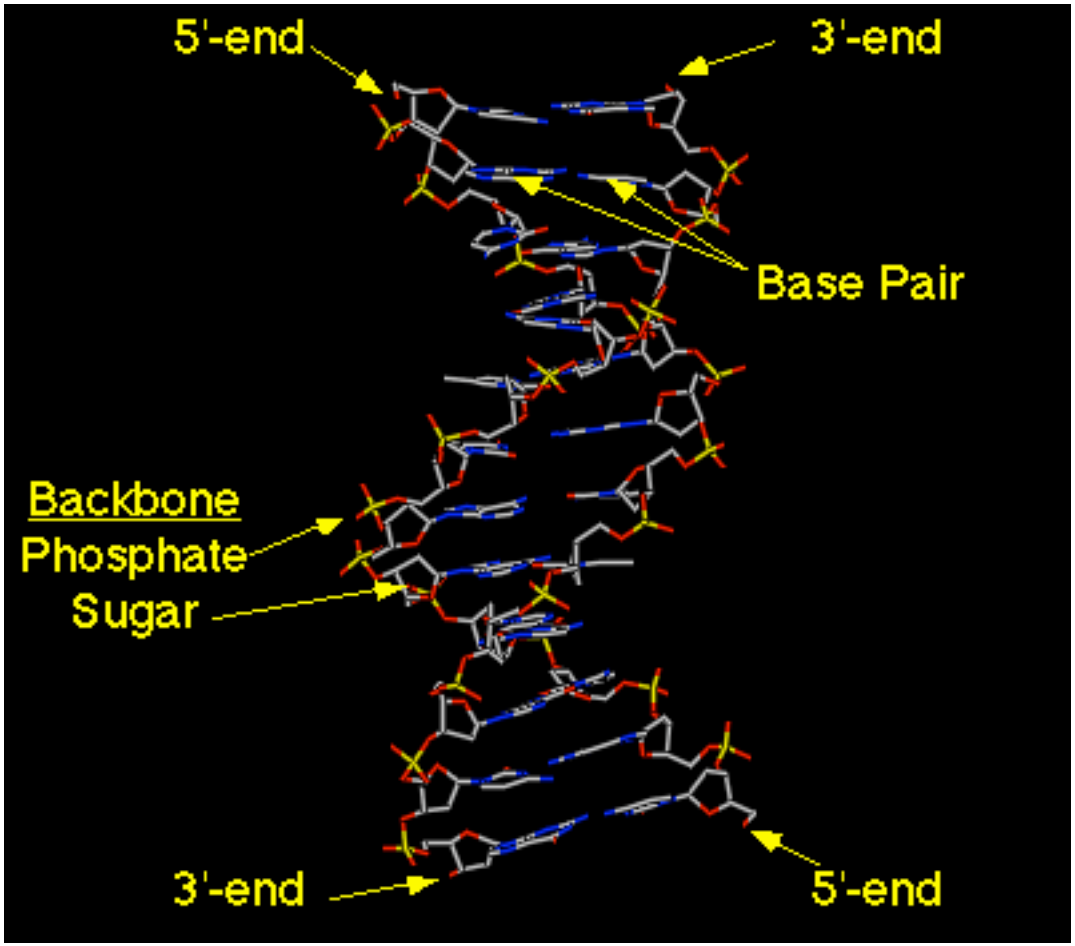
Learning Objectives

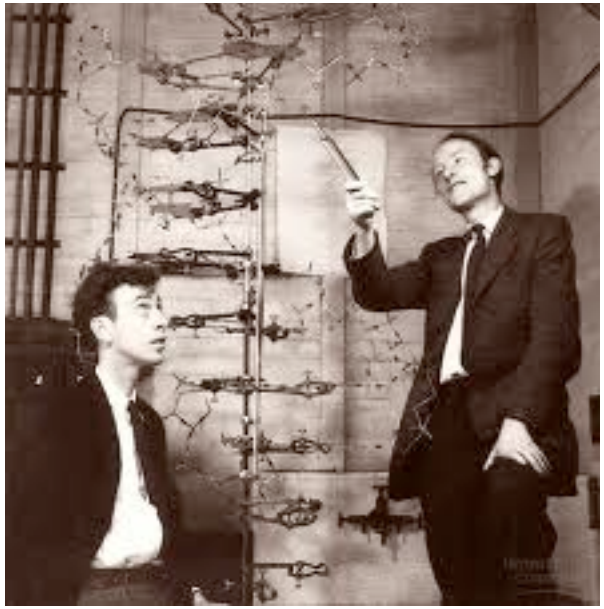
- Outline the basic levels of protein structure.
- Outline key differences between X-ray crystallography and NMR spectroscopy.
- Identify relevant parameters for evaluating the quality of protein structures.

Outline

- The first 3D-structures - DNA & protein
- Protein structure evolution and function
  - Inferring function from structure.
  - Modifying function
- Experimental techniques
  - X-ray crystallography
  - NMR spectroscopy
- Structure validation

DNA - a double helix

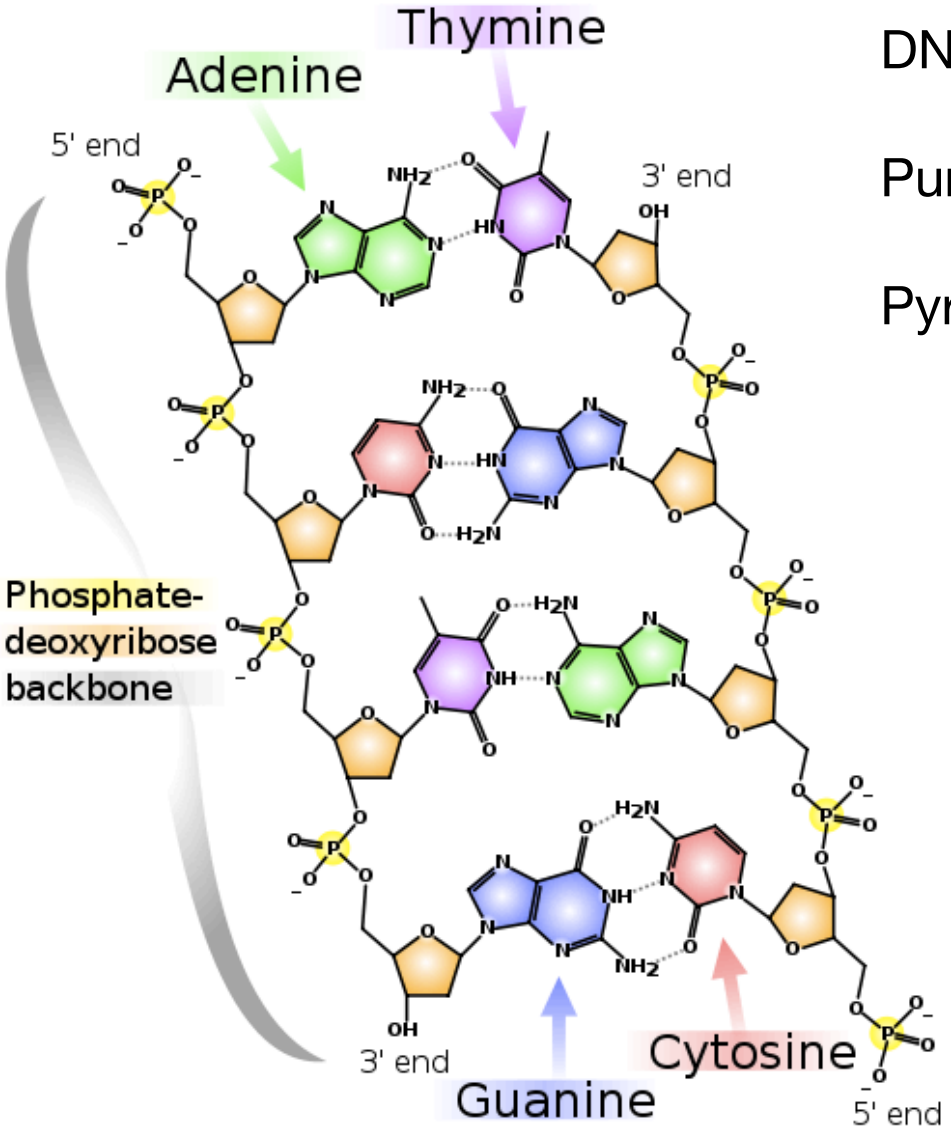




James Watson and Francis Crick with their model of the structure of the DNA molecule, 1953

→ 5' – A T T G C C – 3'
3' – T A A C G G – 5' ←

DNA - a double helix



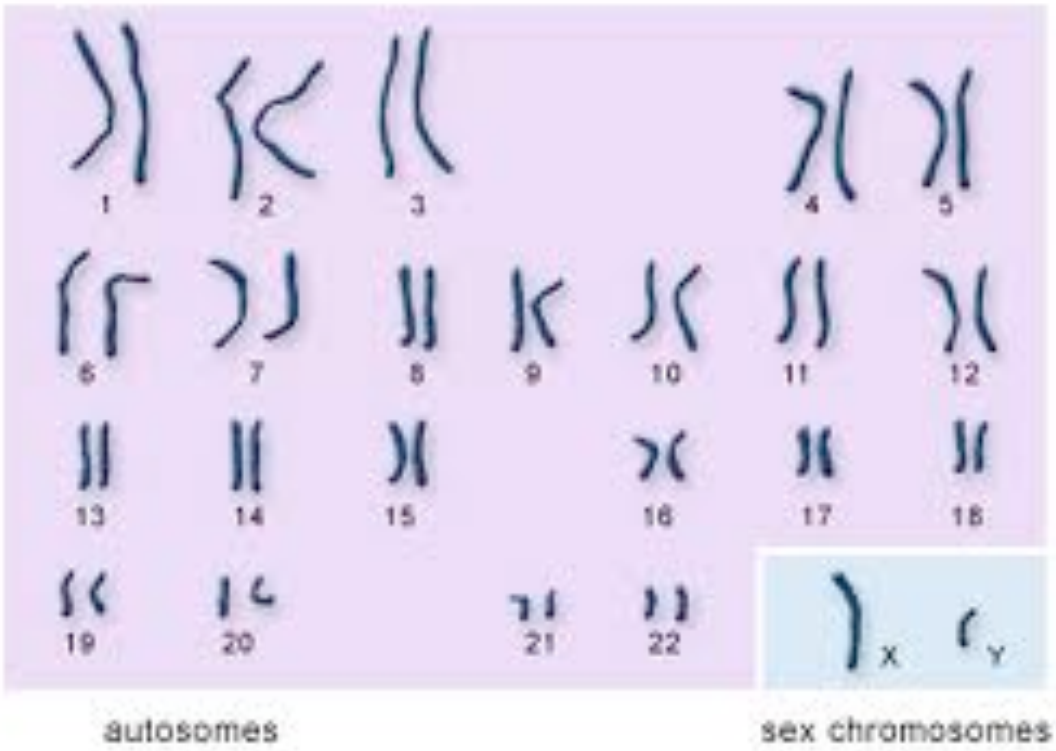
DNA – DeoxyriboNucleic Acid

Purines: A G

Pyrimidines: T C

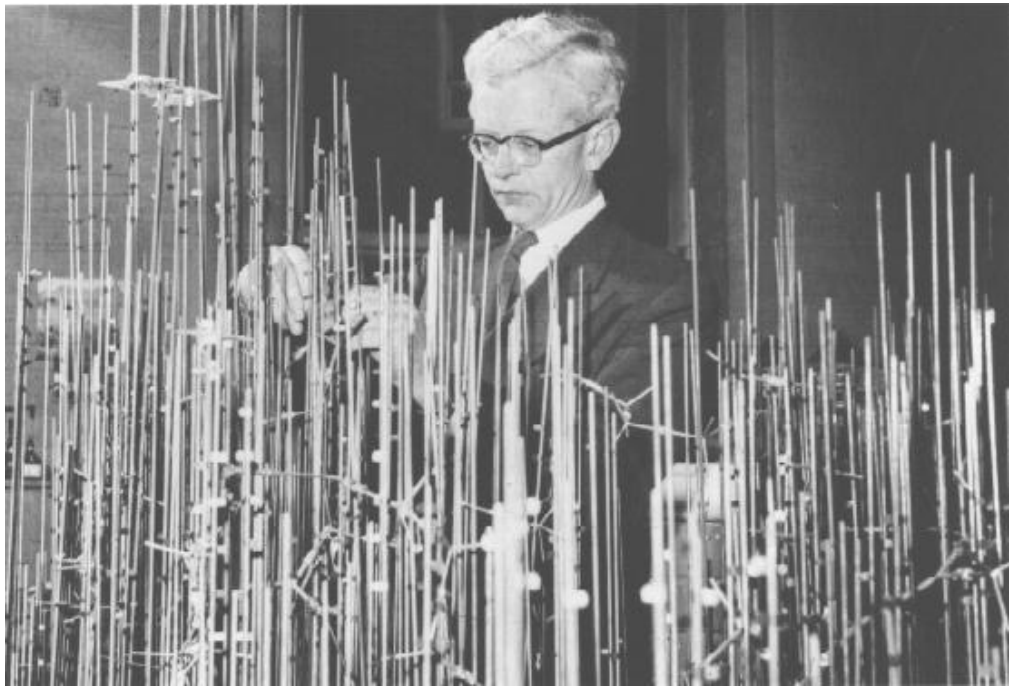
Genes, chromosomes and base pairs

- Genes are located at the chromosomes
- 3.000.000 bp in human genome - diploid => 6.000.000 bp



Once Upon a Time…

"Could the search for ultimate truth really have revealed so hideous and visceral-looking an object?" Max Perutz, 1964, on protein structure



John Kendrew, 1959, with myoglobin model

Why are Protein Structures so Interesting?

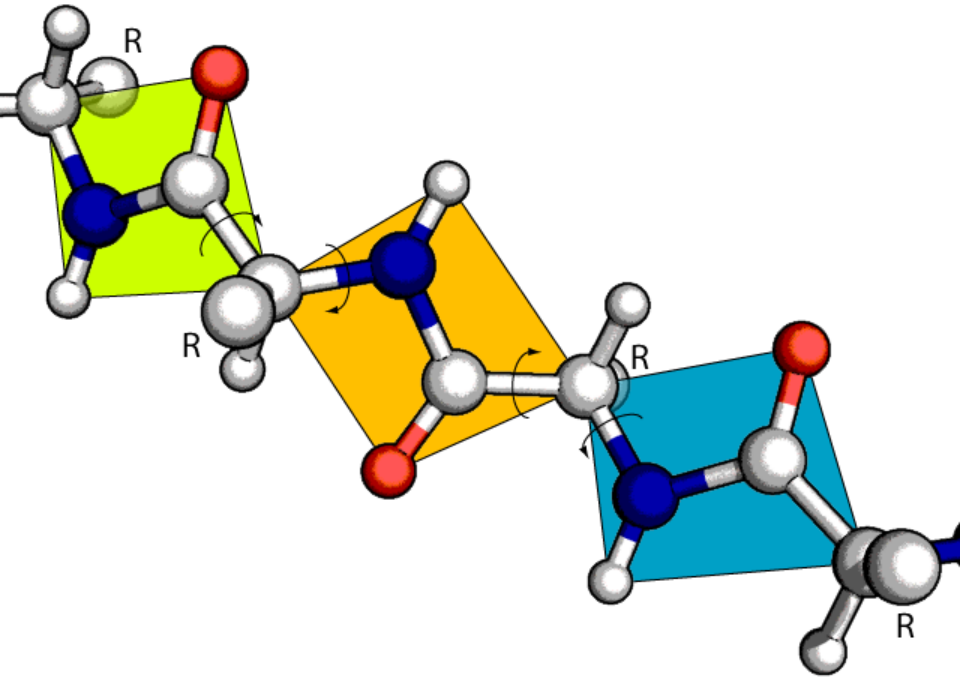
- They provide a detailed picture of interesting biological features: Overall structure and amino acids involved in active site.

- They aid in rational drug design and protein engineering.

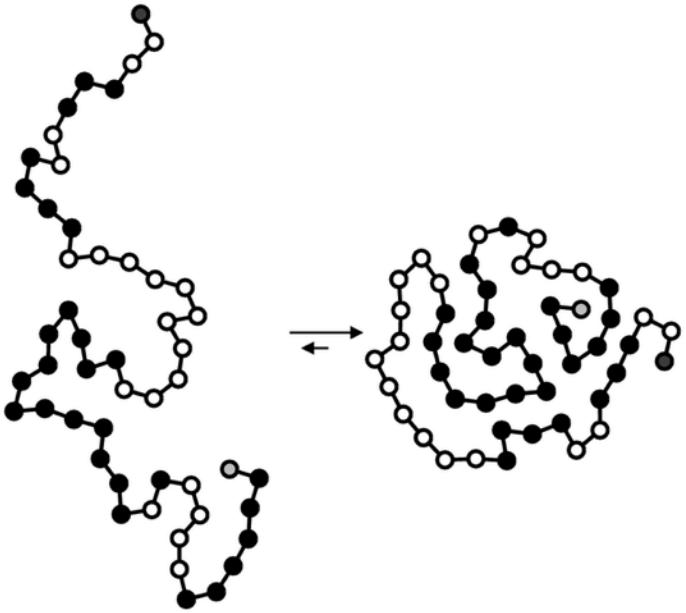
- They can elucidate evolutionary relationships undetectable by sequence comparisons.

Proteins Are Polypeptides

- A polypeptide chain



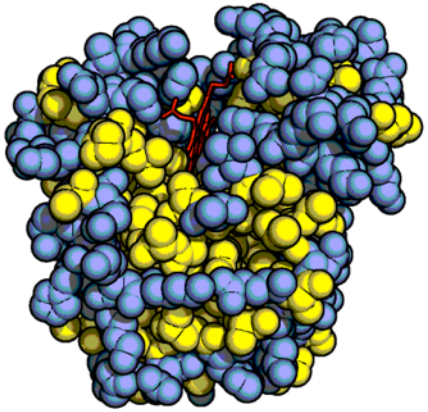
- Hydrophobic collapse



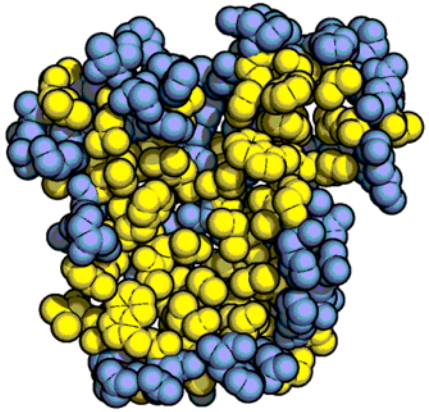
Protein Folding

- Hydrophobic collapse
  - Hydrophobic residues cluster to “escape” interactions with water.

Myoglobin



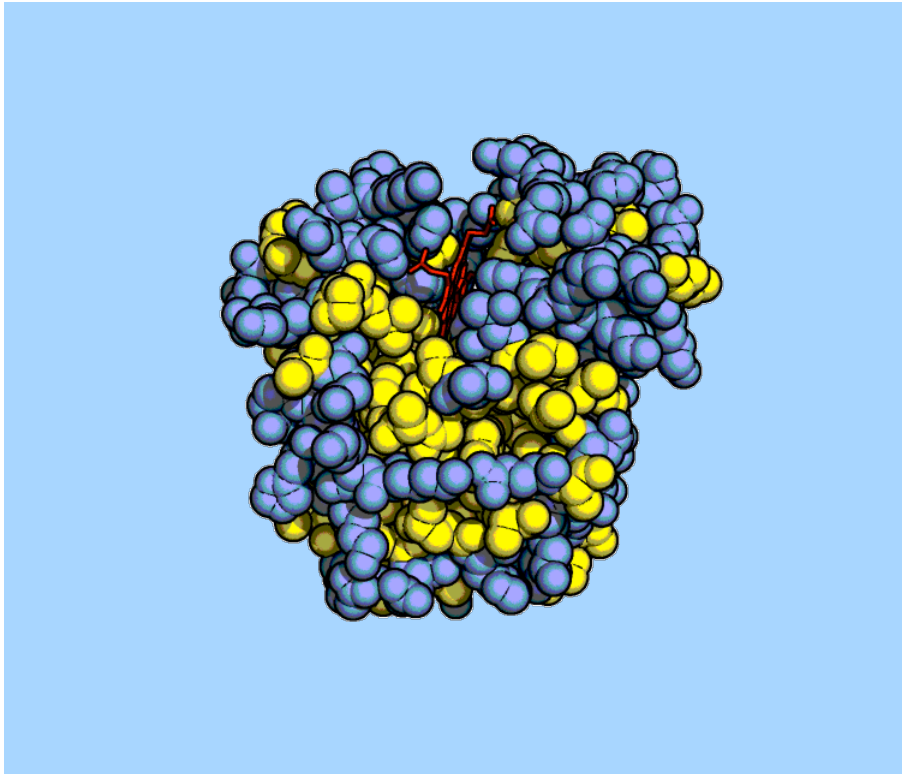
Surface



Interior

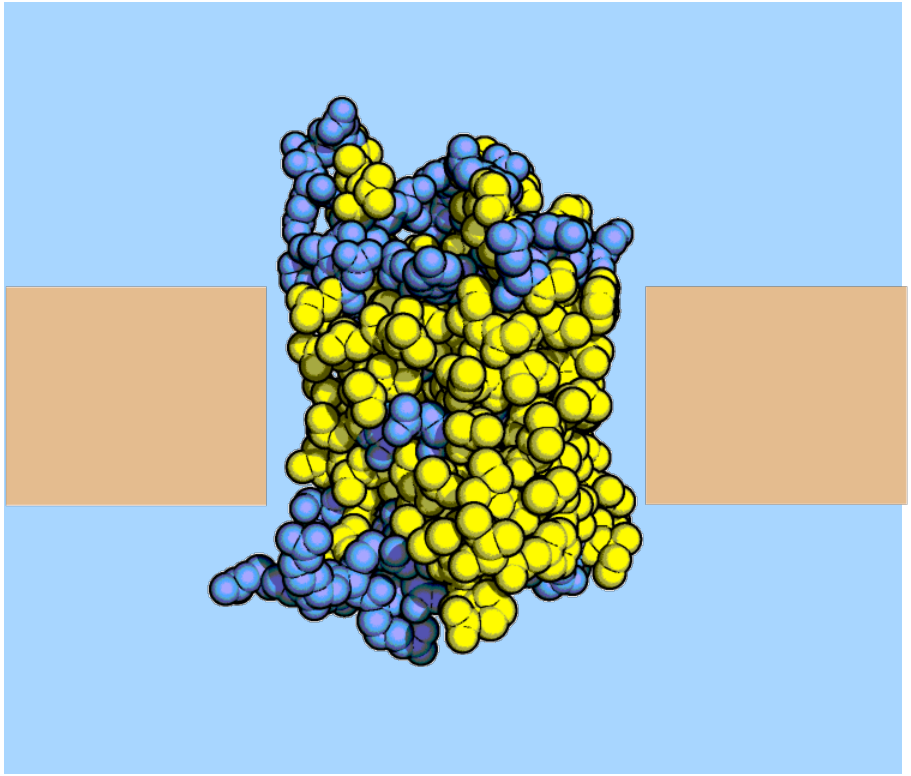
Hydrophobic vs. Hydrophilic

- Globular protein (in solution)



Myoglobin

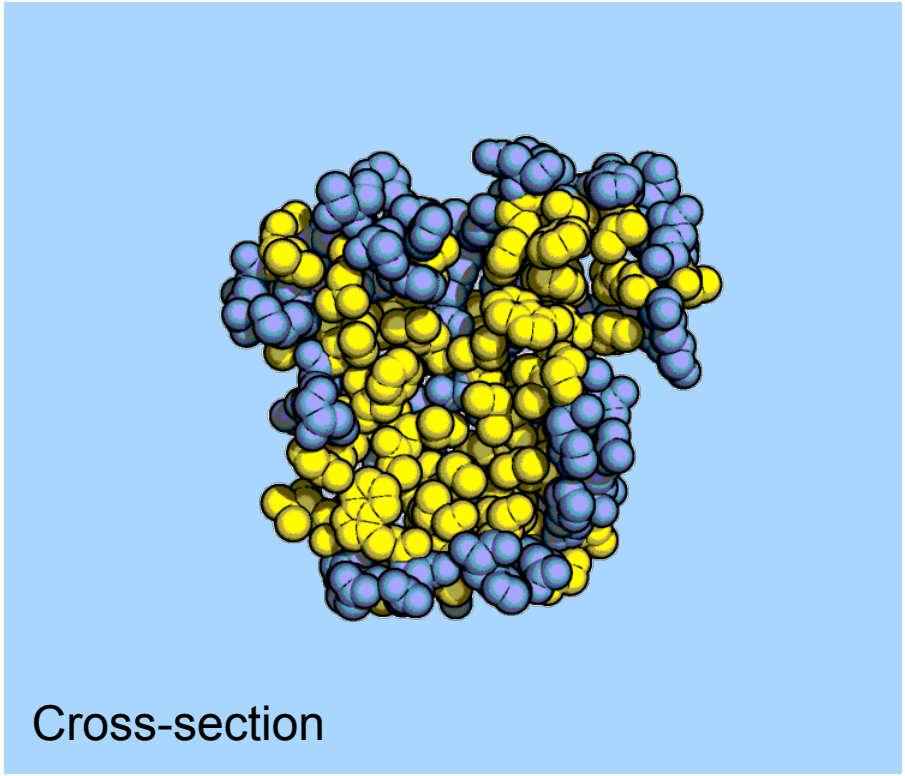
- Membrane protein (in membrane)



Aquaporin

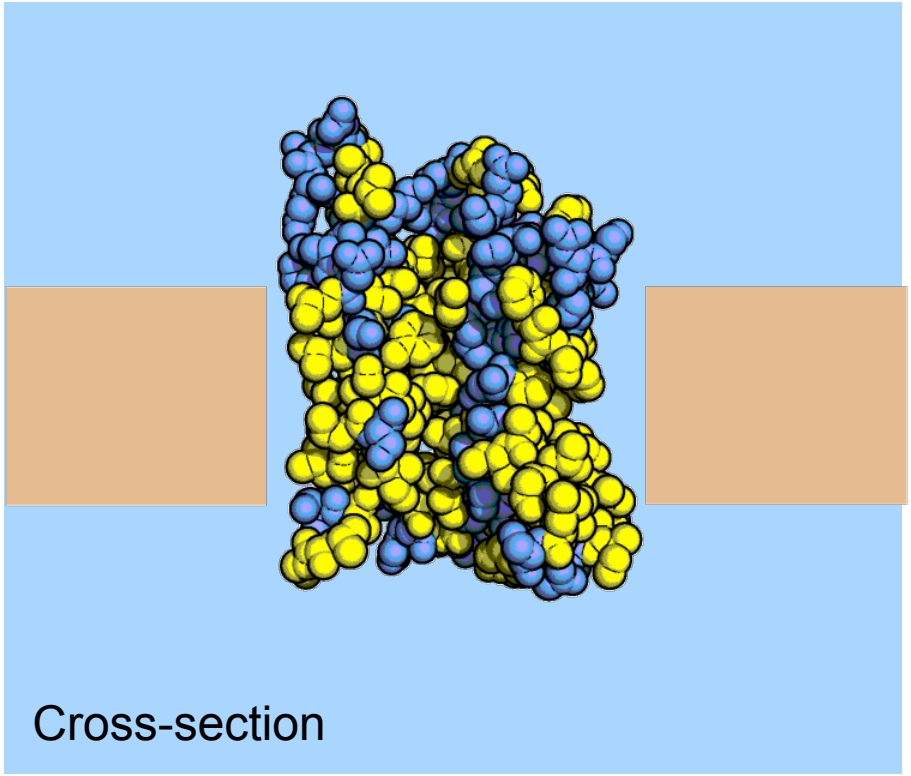
Hydrophobic vs. Hydrophilic

- Globular protein (in solution)



Myoglobin

- Membrane protein (in membrane)



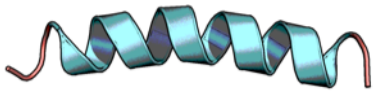
Aquaporin

Structure Levels

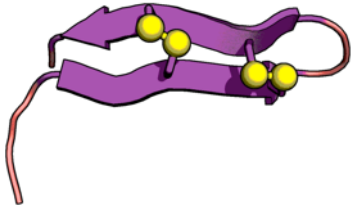
- Primary structure = Sequence

MSSVLLGHIKKLEMGHS…

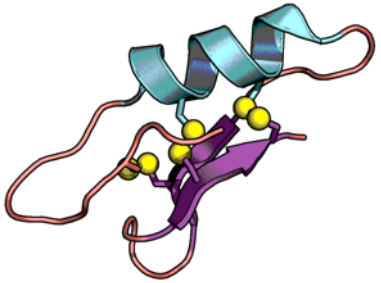
- Secondary Structure = Helix, sheets/strands, loops & turns



- Structural Motif = Small, recurrent arrangement of secondary structure, e.g.
  - Helix-loop-helix
  - Beta hairpins
  - EF hand (calcium binding motif)
  - Etc.



- Tertiary structure = Arrangement of Secondary structure elements



Quaternary Structure

- Assembly of several polypeptide chains into a protein complex.

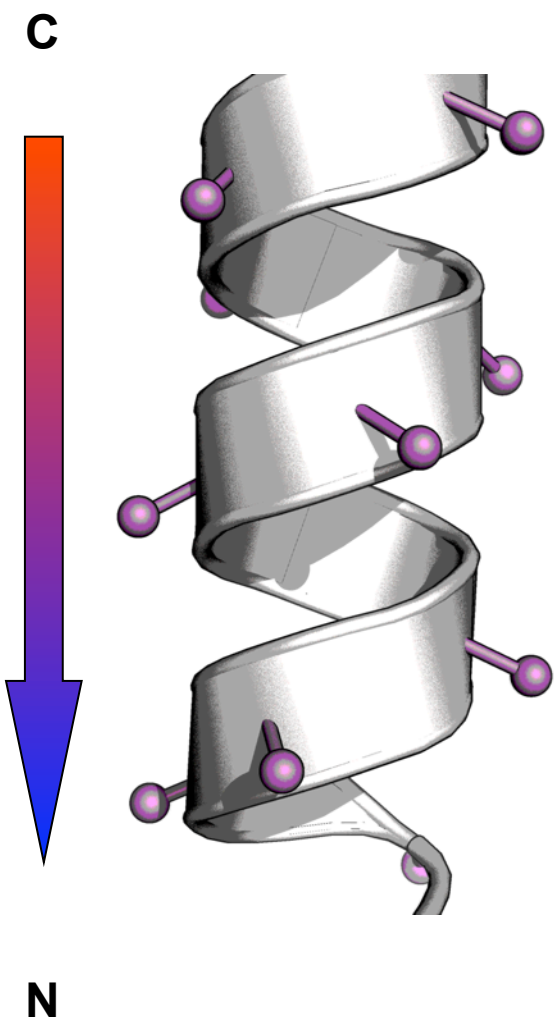
- Hemoglobin is functionally a tetramer composed of two different chains α and β. It is therefore a hetero-tetramer.

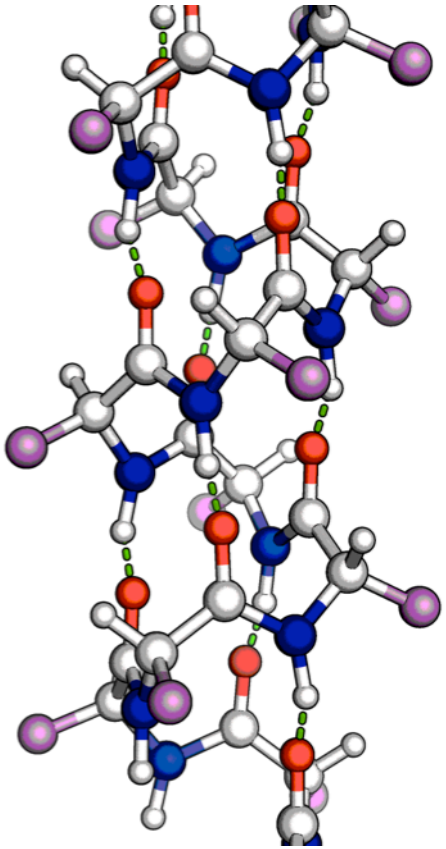
- Myoglobin

- Hemoglobin

Characteristics of Helices

- Aligned peptide units → Dipolar moment
- Ion/ligand binding
- Secondary and quaternary structure packing
- Capping residues
- The α helix (i→i +4)
- Other helix types! (3_{10}, π)

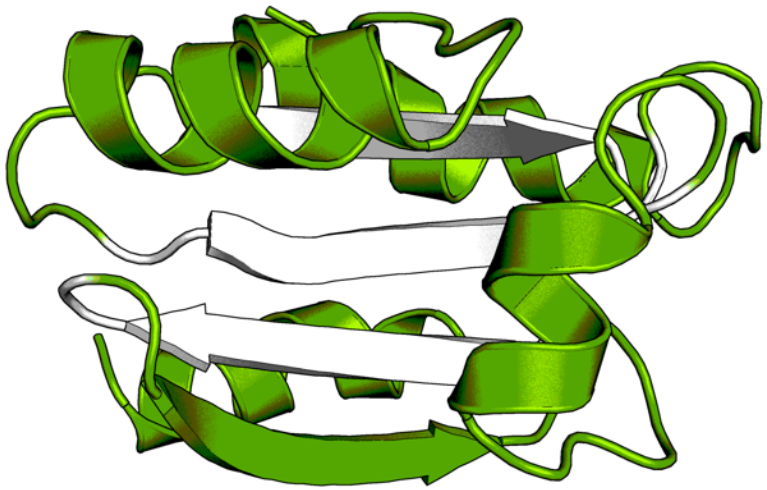




β-Sheets

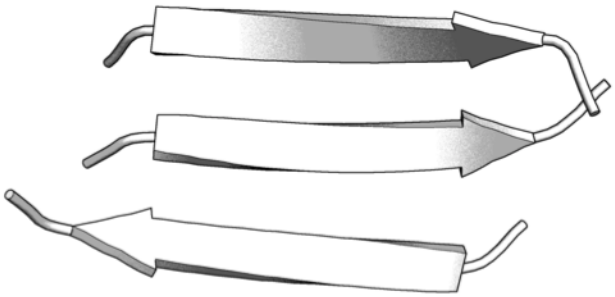
- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

Thioredoxin



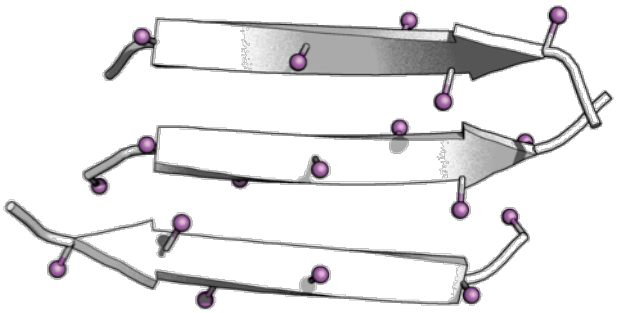
β-Sheets

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist



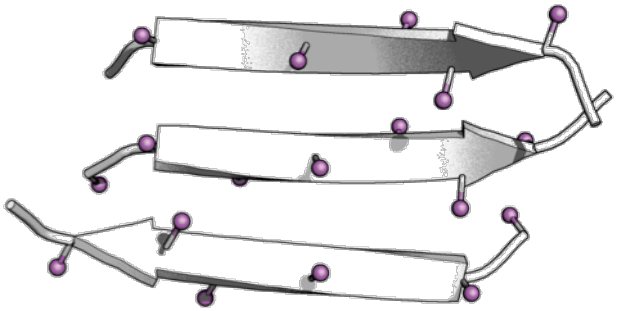
β-Sheets

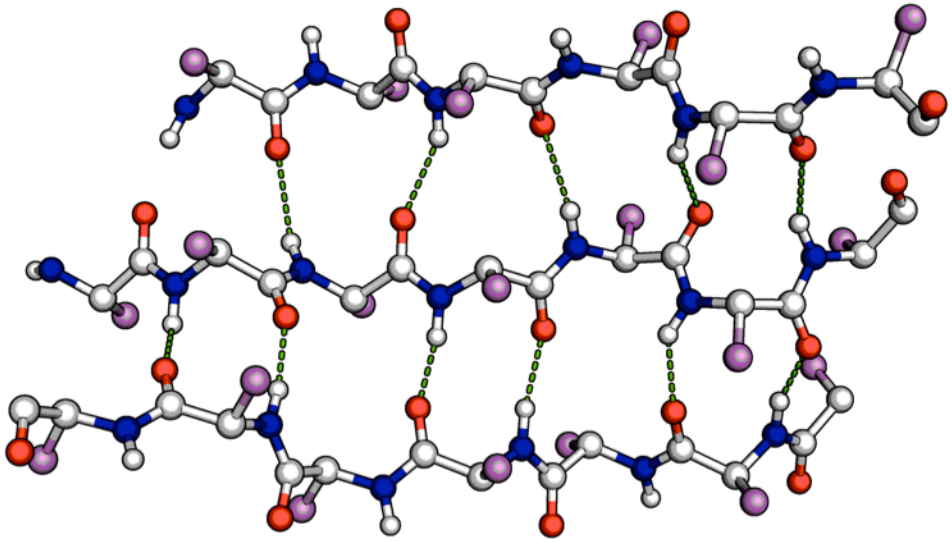
- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist



β-Sheets

- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist



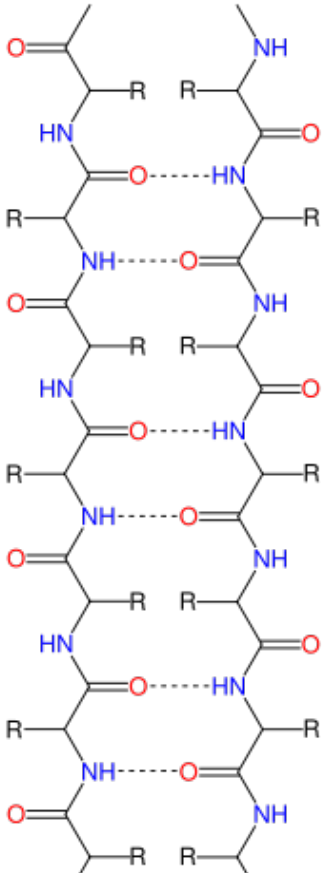


β-Sheets

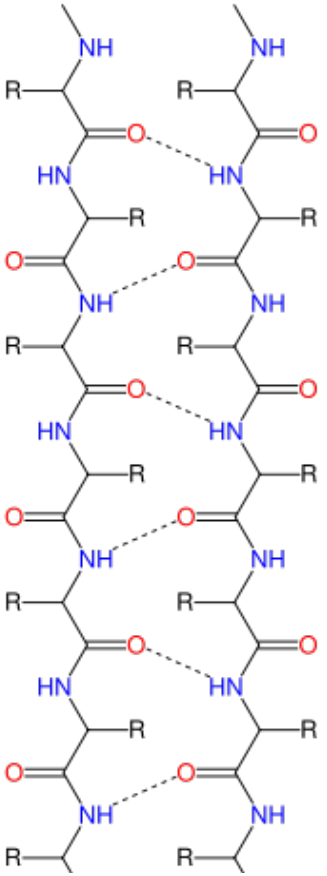
- Multiple strands → sheet
  - Parallel vs. antiparallel
  - Twist

- Strand interactions are **non-local**

Antiparallel

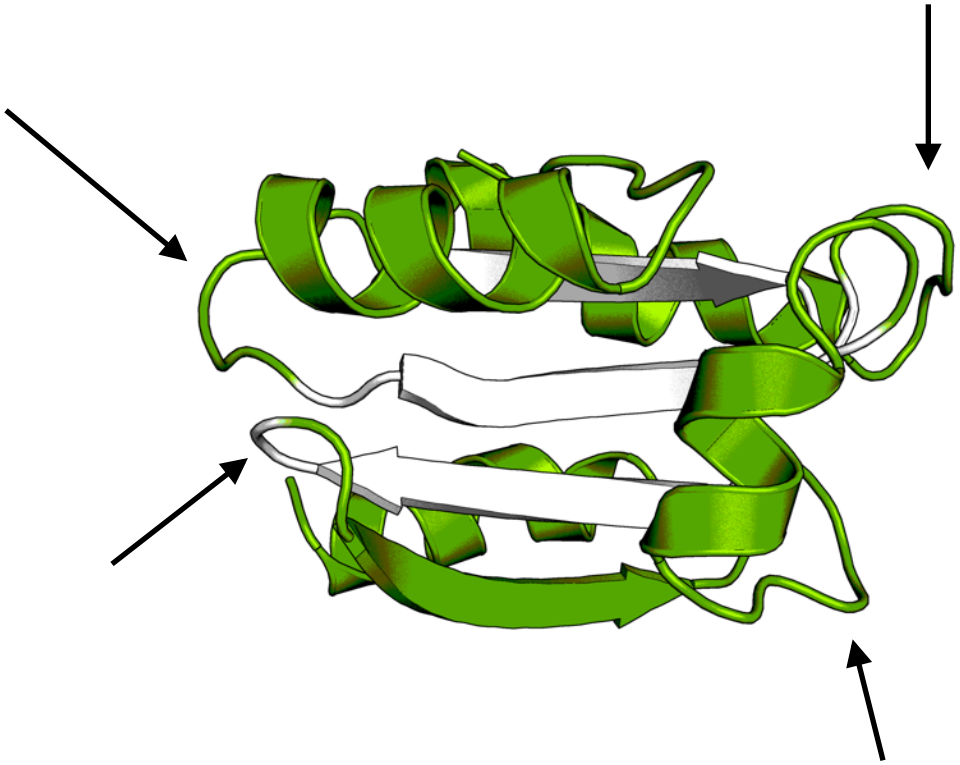


Parallel



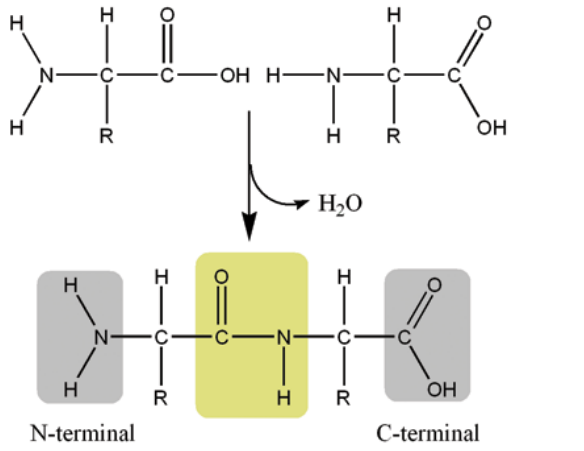
Turns, Loops & Bends

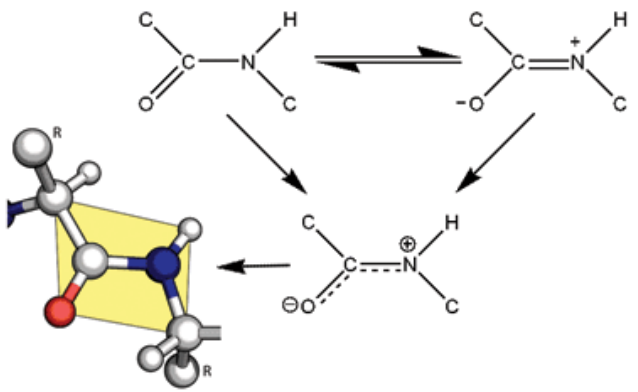
- Connecting secondary structure (SS) elements
- Often at the surface
- More flexible than SS



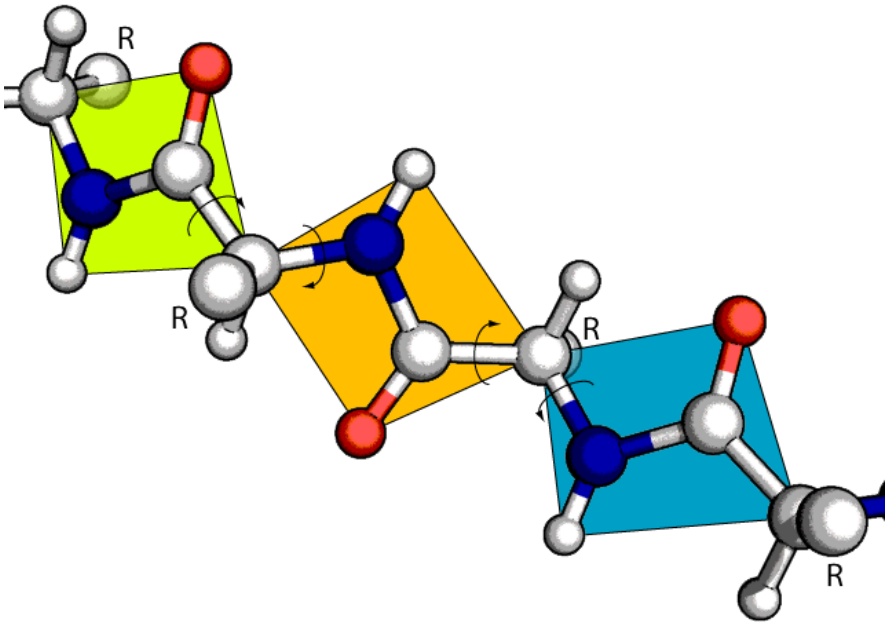
Proteins Are Polypeptides

- The peptide bond



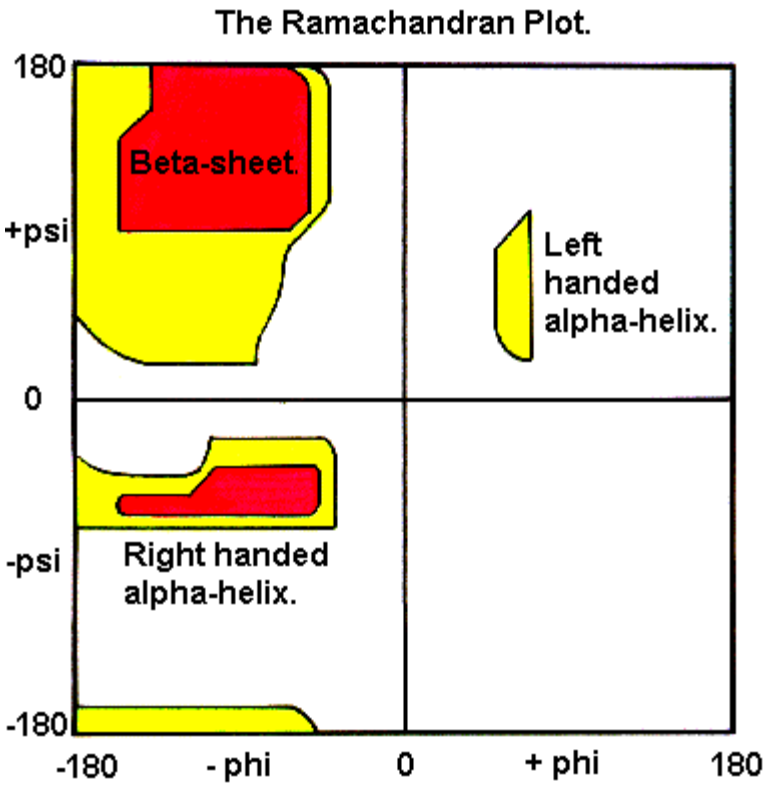


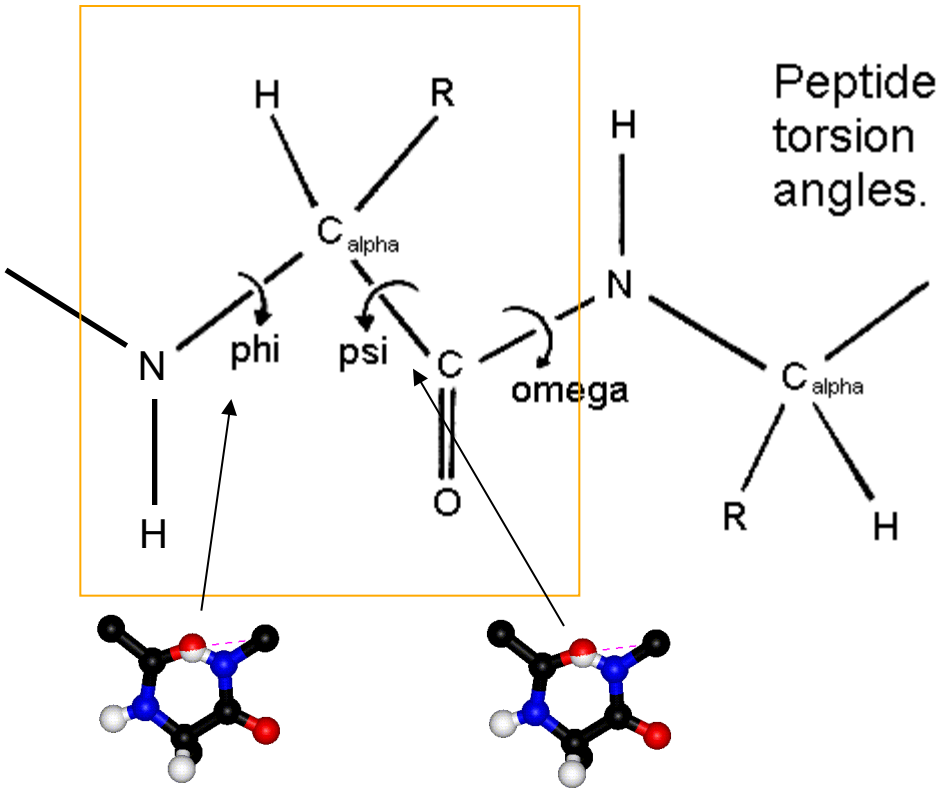
- A polypeptide chain



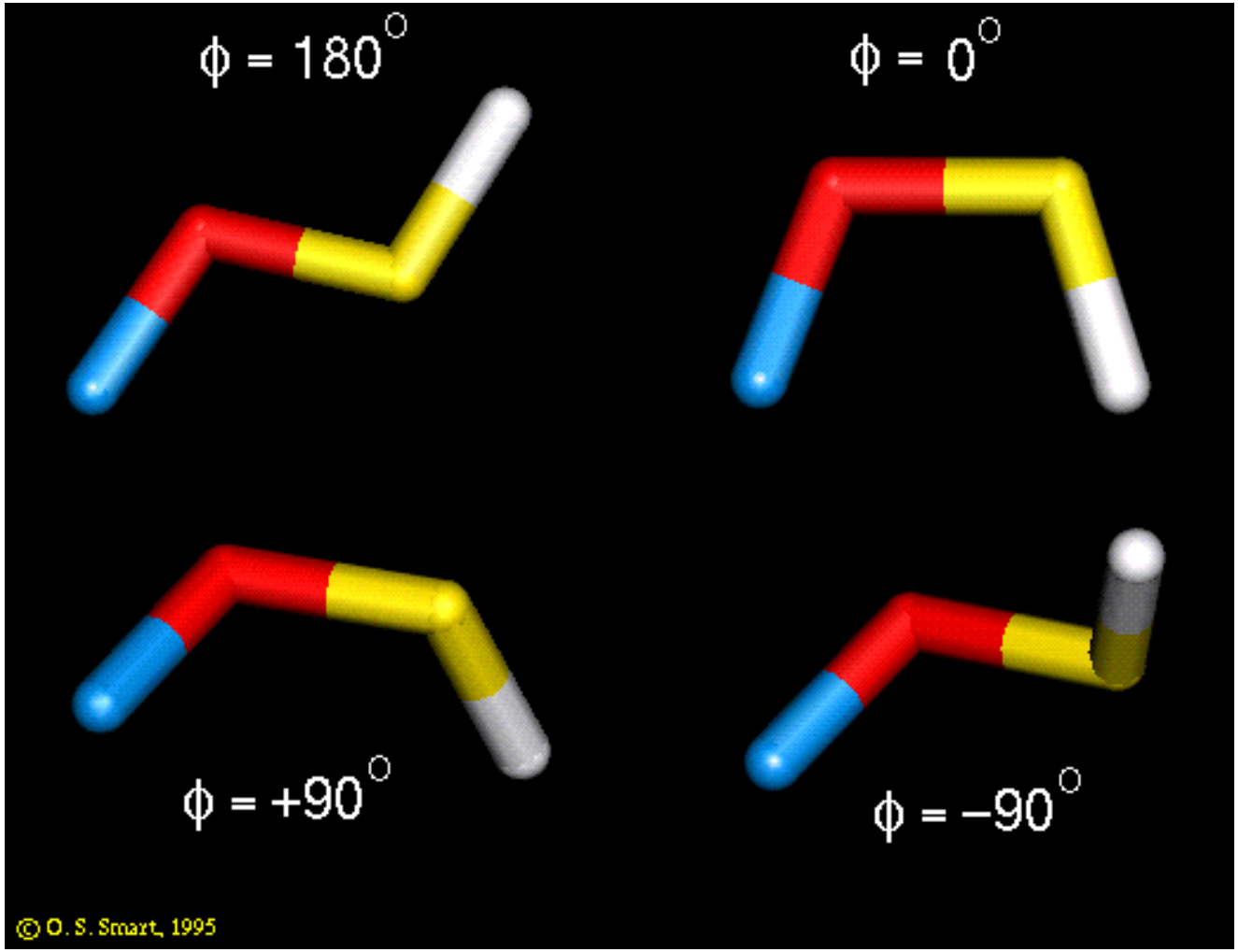
Ramachandran Plot

- Allowed backbone torsion angles in proteins

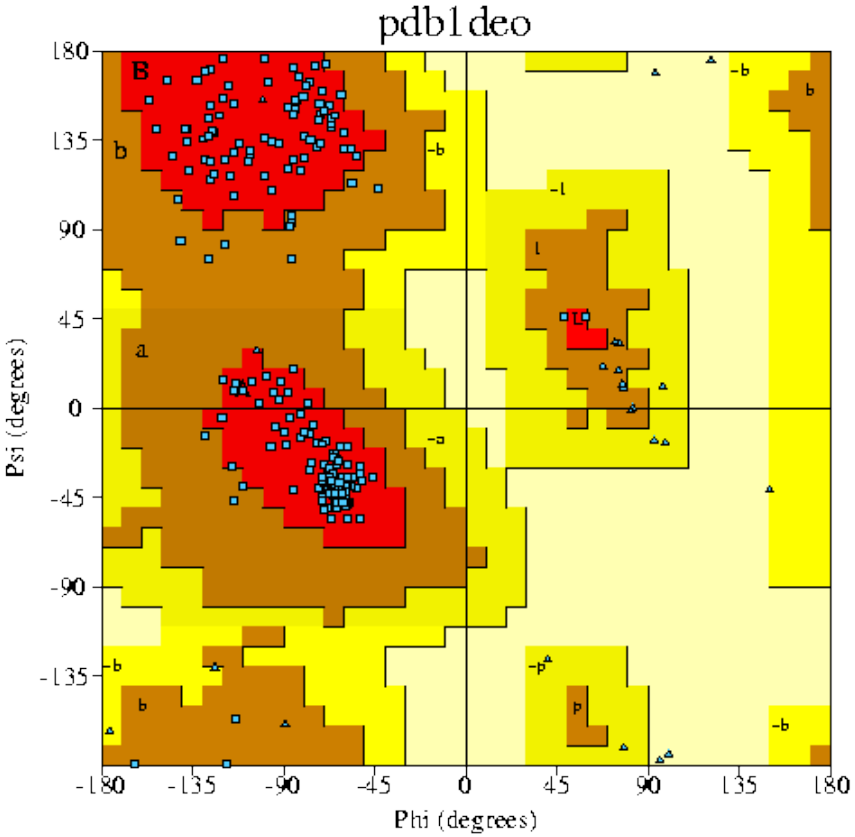


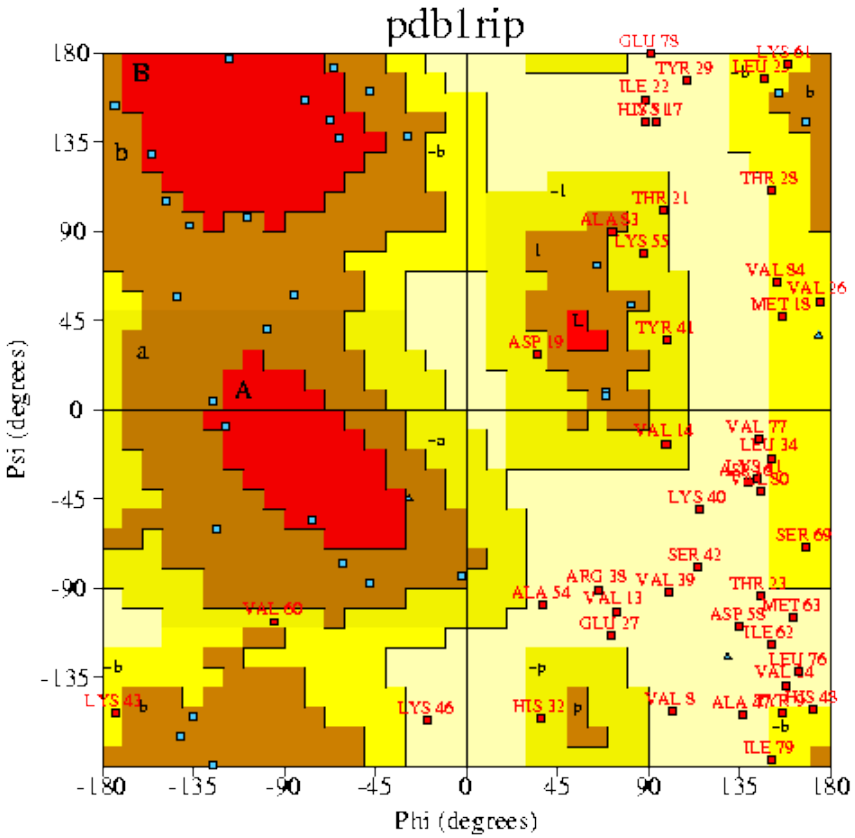


Torsion Angles



Ramachandran Plots

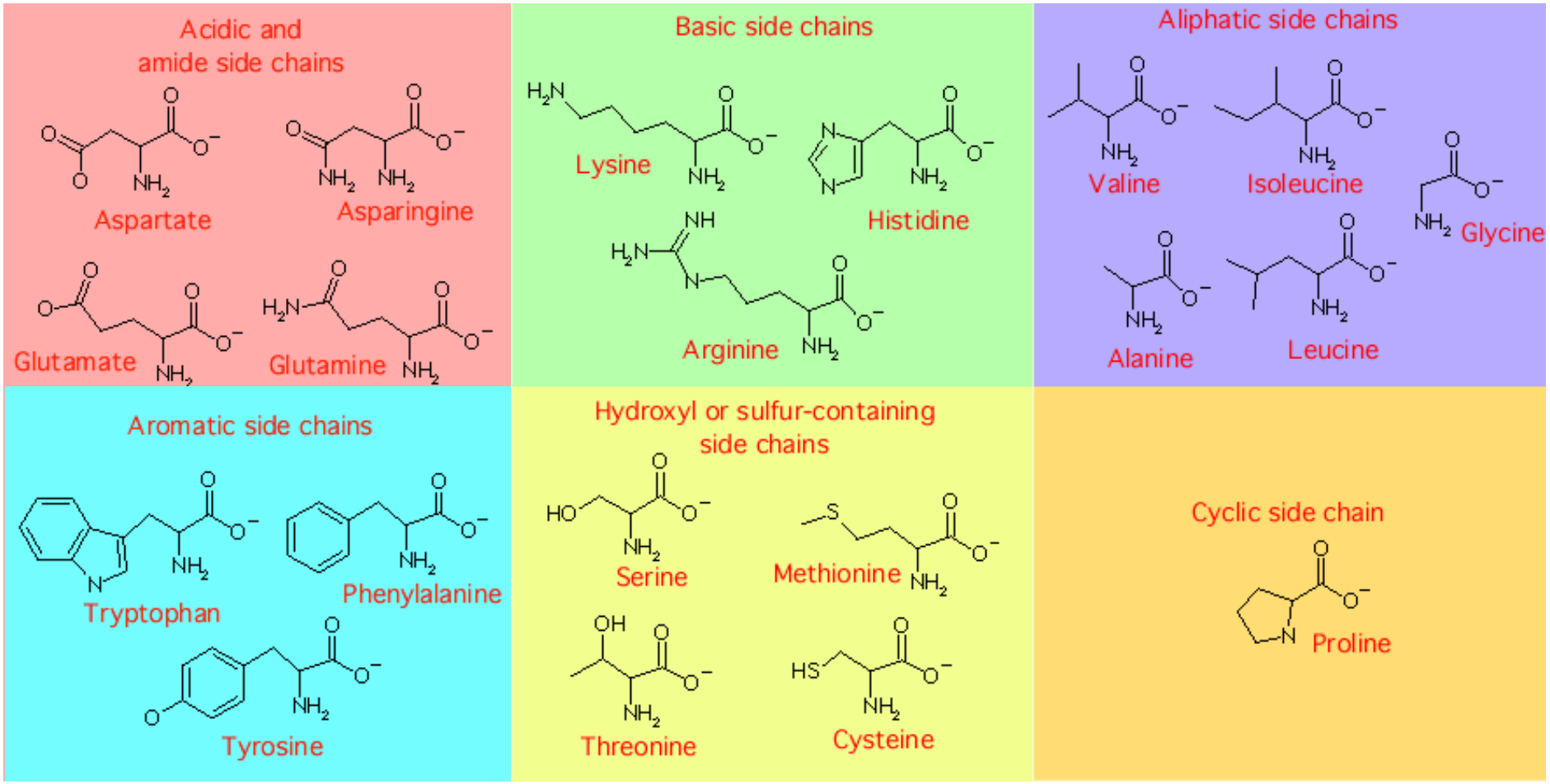




X-ray structure – good data.

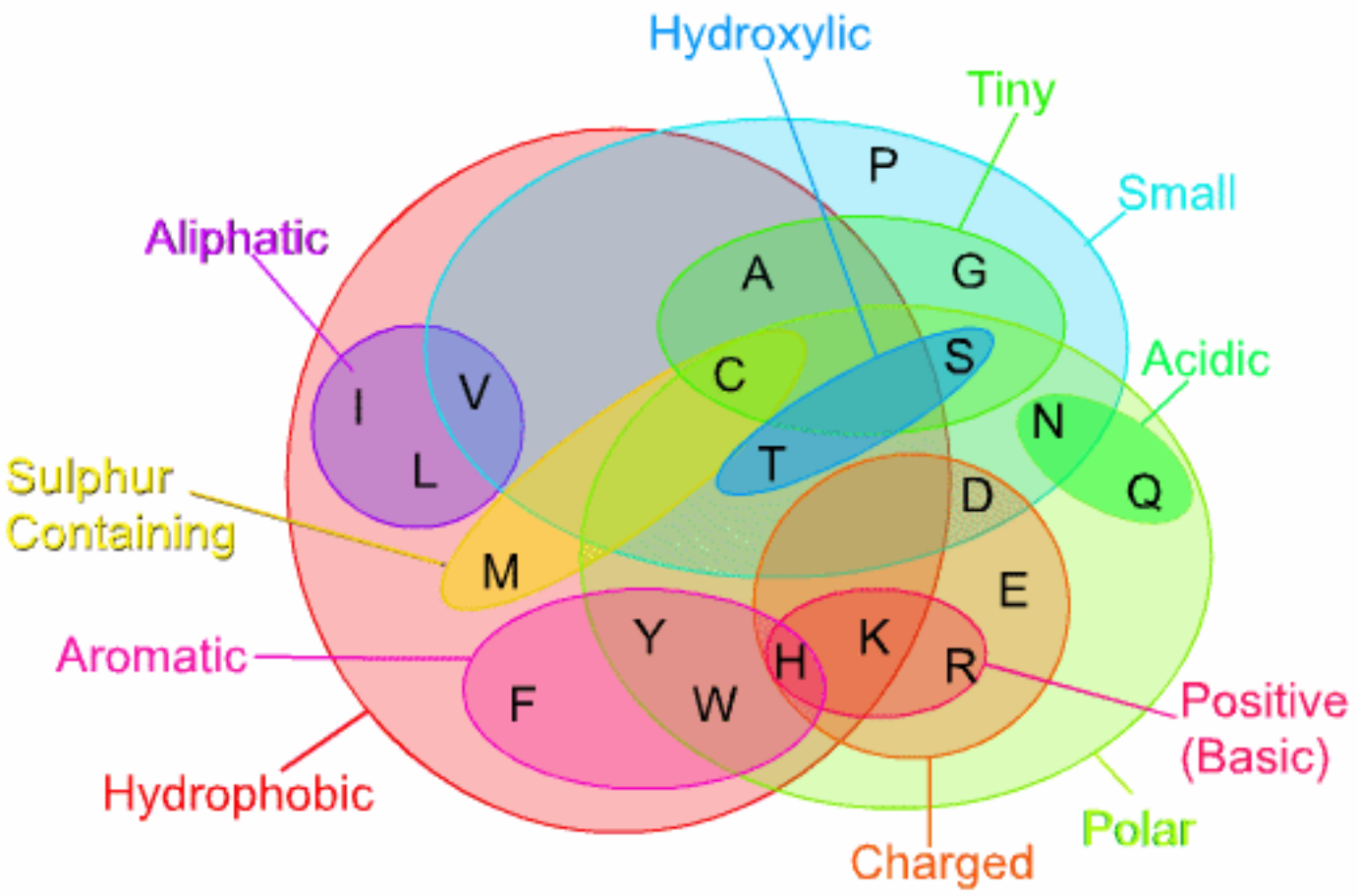
NMR structure – low quality data…

The Amino Acids



http://www.ch.cam.ac.uk/magnus/molecules/amino/

Grouping Amino Acids



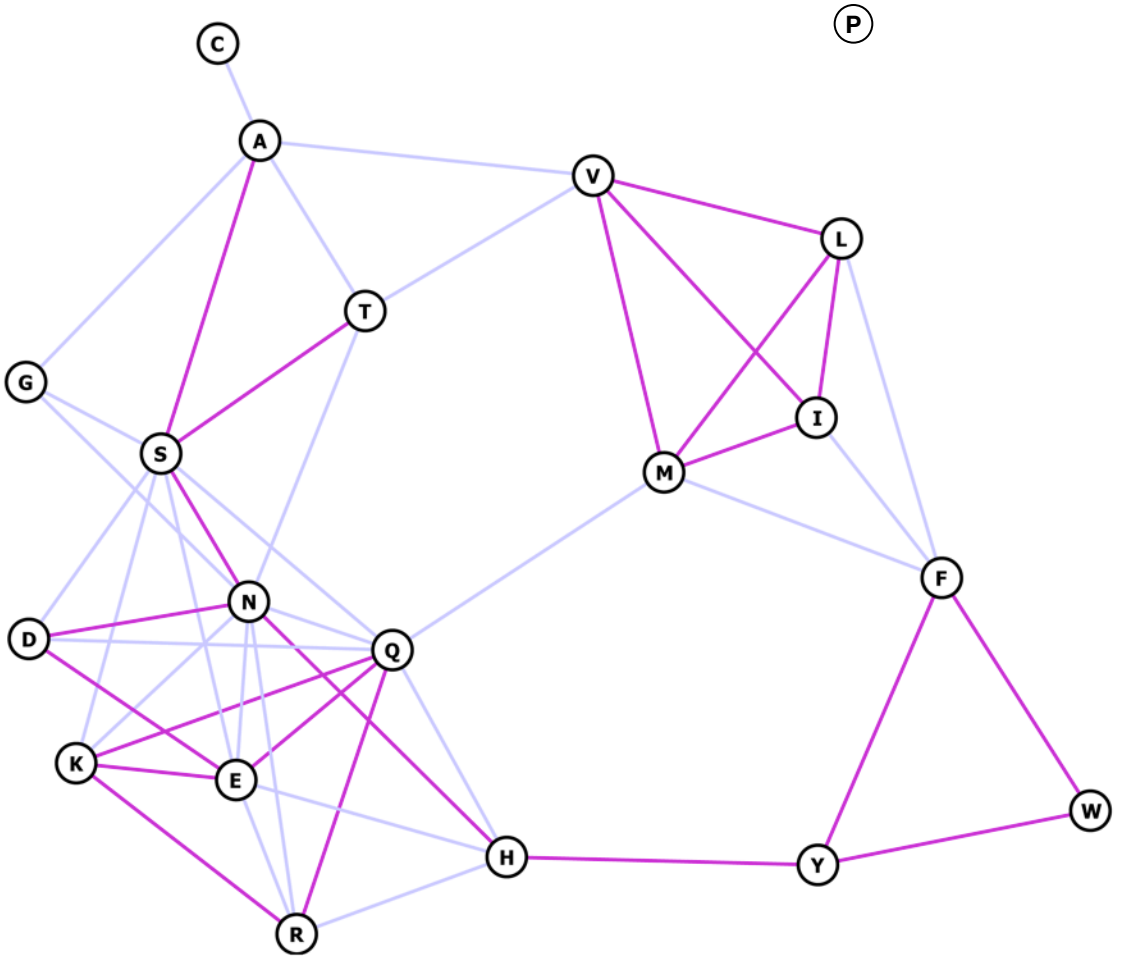
Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

http://www.dreamingintechnicolor.com/InfoAndIdeas/AminoAcids.gif

The Evolution Way

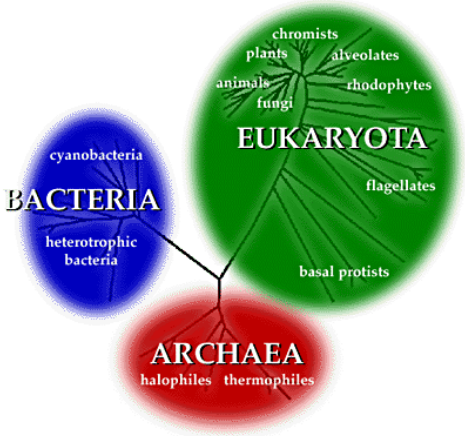
- Based on Blosum62 matrix
- Measure of evolutionary substitution probability



Structure & Evolution

- The structure of a protein is uniquely determined by its amino acid sequence
  (but sequence is sometimes not enough):
  - prions
  - pH, ions, cofactors, chaperones

- In evolution structure is conserved much longer than both function and sequence.
  - Structure > Function > Sequence



Form vs. Function

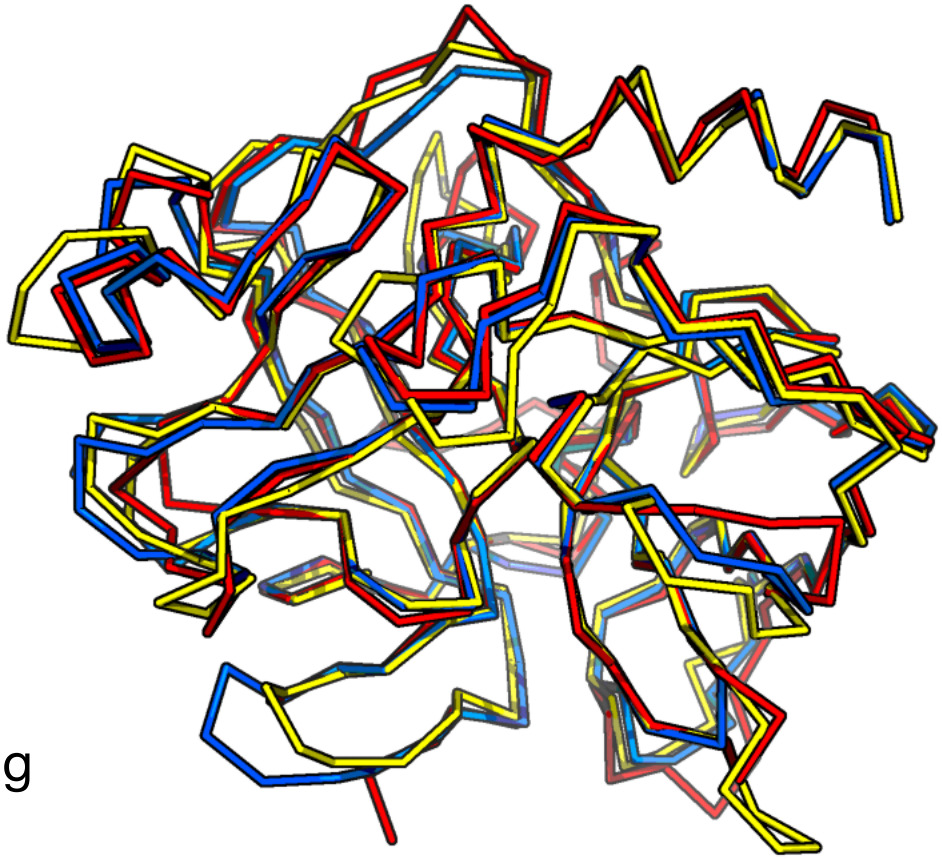
- Divergent evolution
  - Common ancestor
  - New function
- Convergent evolution
  - Different ancestor
  - Same function

Divergent

Convergent

Sequence vs. Function – I

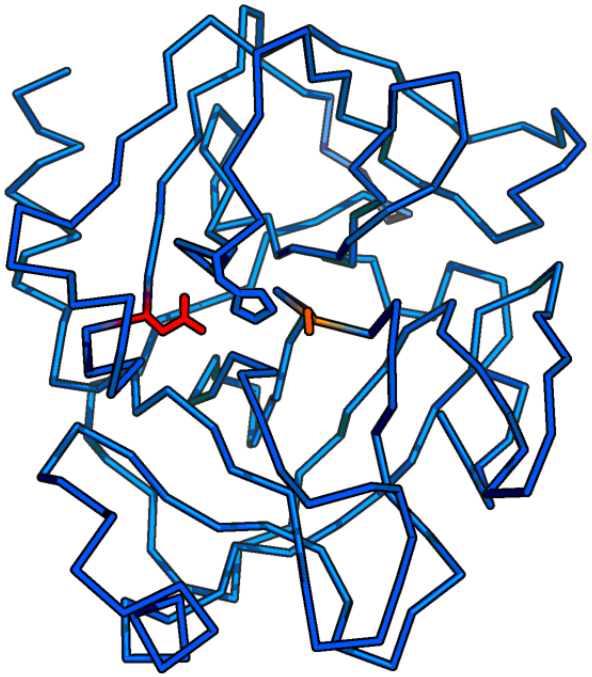
- Trypsin
  - positive
- Chymotrypsin
  - large hydrophobic
- Elastase
  - Small hydrophobic



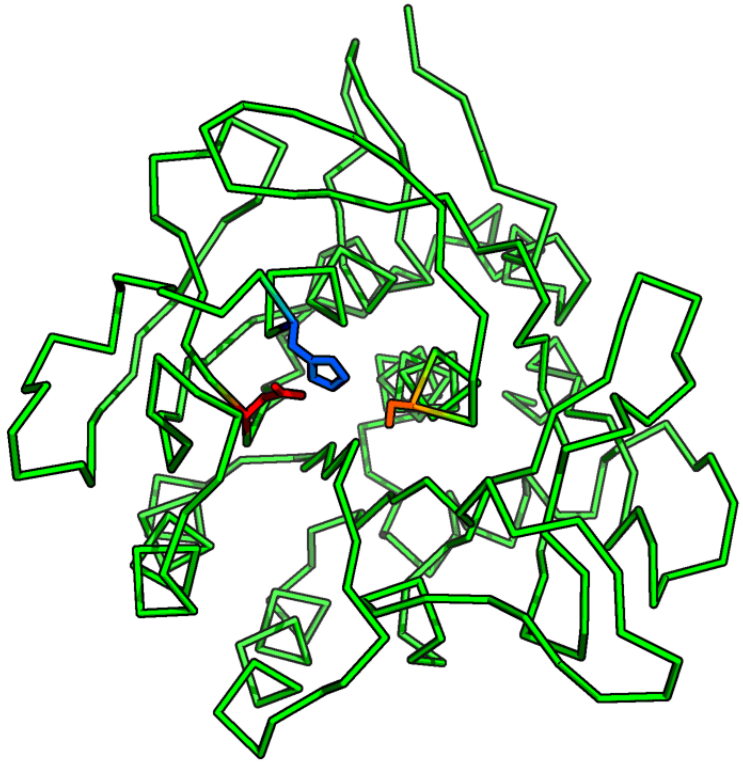
- Divergent evolution
  - Same fold
  - Different specificities
  - Small changes in binding pocket

Sequence vs. Function – II

- Trypsin



- Subtilisin

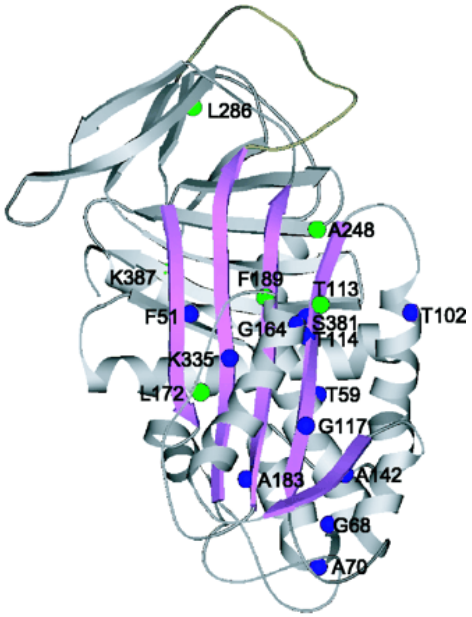


Convergent evolution

Engineering & Design

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

- Protein engineering
  - Overpacking
  - Buried polar groups
  - Cavities

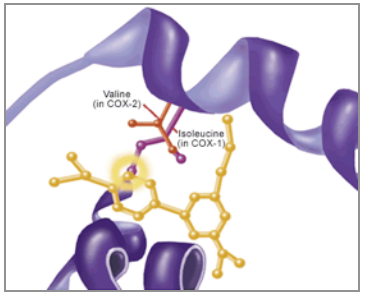


Im, Ryu & Yu (2004), *Engineering thermostability in serine protease inhibitors*, PEDS, 17, 325-331.

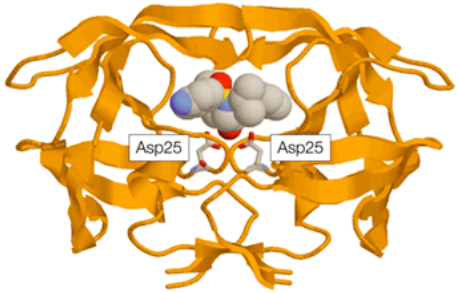
- Drug design
  - Target specificity/selectivity
  - Function
  - Mutations

COX-1/COX-2
- Arthritis
- Designed to prevent drug side effects

http://publications.nigms.nih.gov/structlife/chapter4.html



HIV protease



Blundell et al. (2002), *High-throughput crystallography for lead discovery in drug design*, Nature Reviews Drug Discovery 1, 45-54.

Engineering & Design

http://www.biosustain.dtu.dk/

Experimental Methods

Crystallography
&
NMR spectroscopy

Methods for Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR)
- Modelling techniques

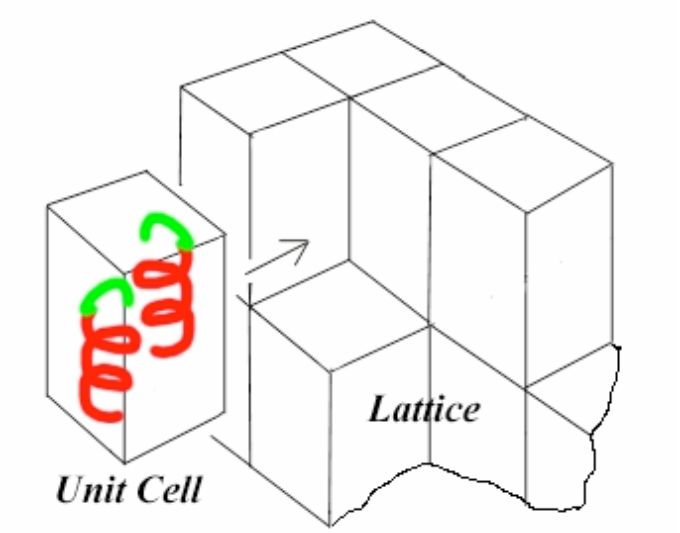
- More exotic techniques
  - Cryo electron microscopy (Cryo EM)
  - Small angle X-ray scattering (SAXS)
  - Neutron scattering

X-ray Crystallography

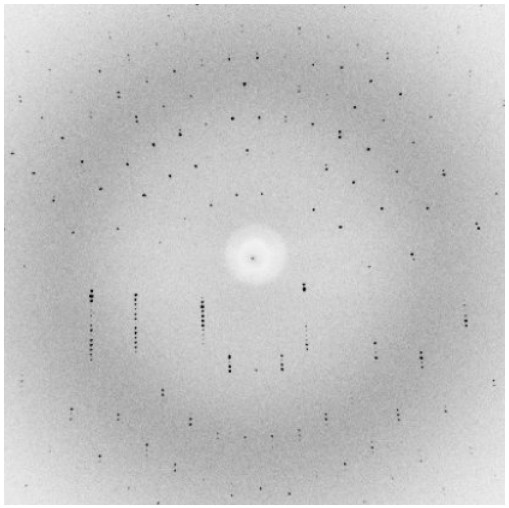
- No size limitation.
- Protein molecules are ”stuck” in a crystal lattice.
- Some proteins seem to be uncrystallizable.
- Slow.

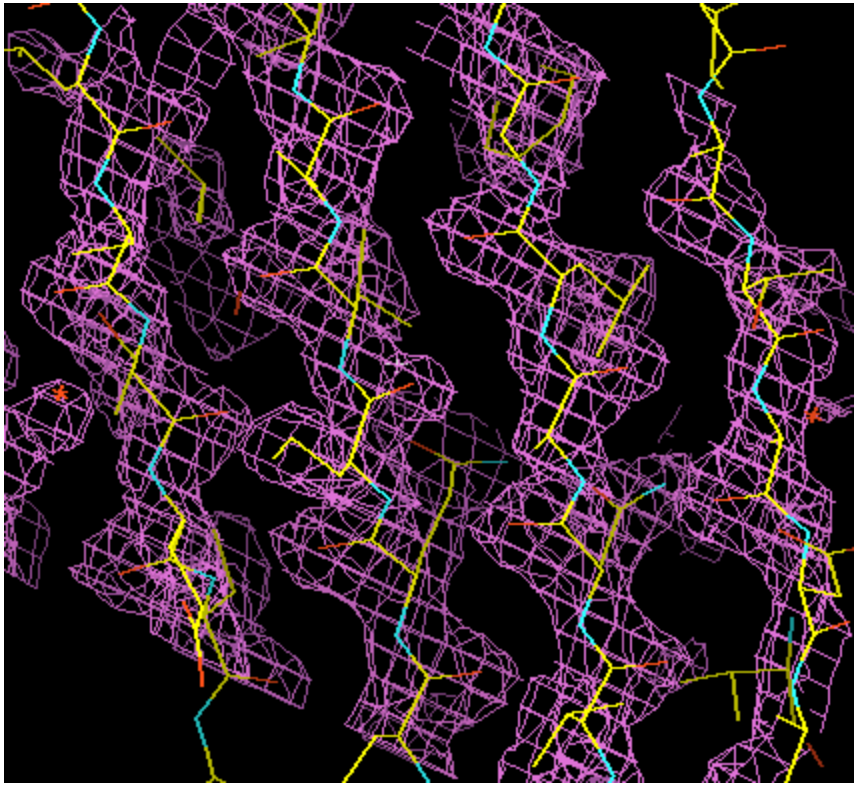
- Especially suited for studying structural details.

- Lattice and unit cell



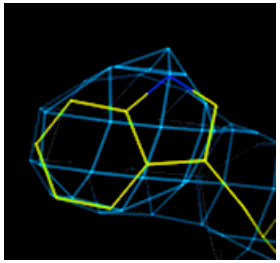
X-rays



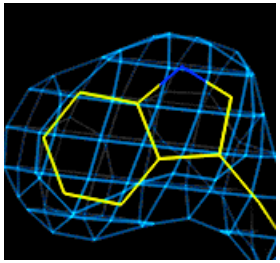


Fourier transform

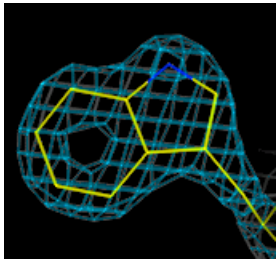
The Importance of Resolution



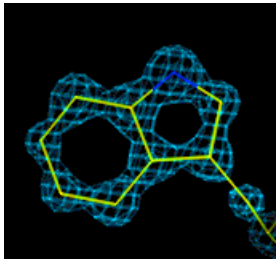
4 Å



3 Å



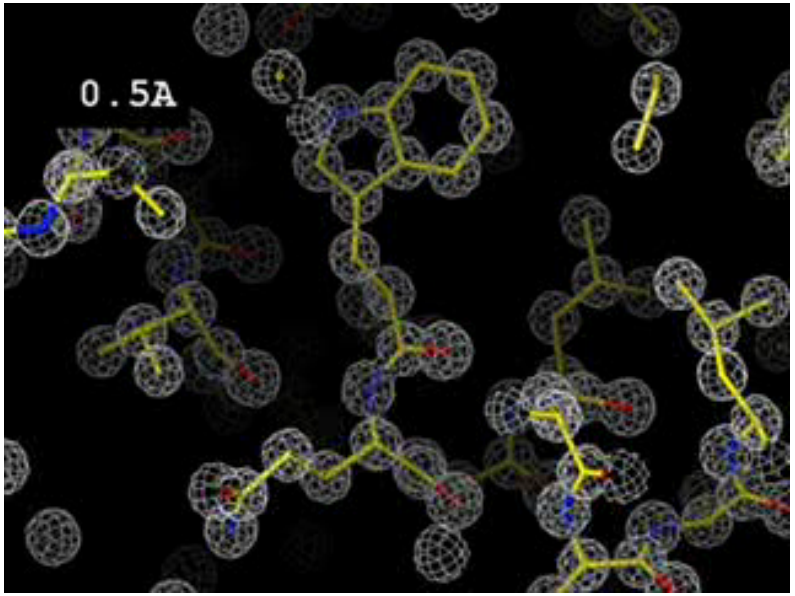
2 Å



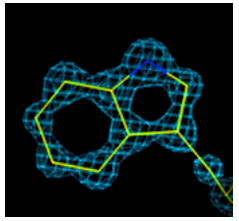
1 Å

low

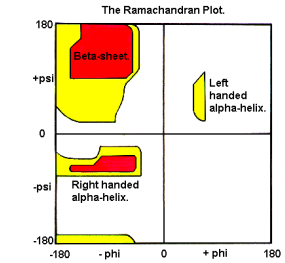
high



Key Parameters

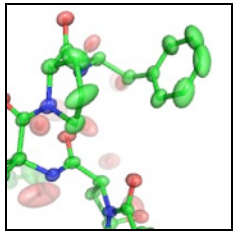


- Resolution
- R values
  - Agreement between data and model.
  - Usually between 0.15 and 0.25, should not exceed 0.30.
    - $R + 0.05 > R_{free} > R$.



- Ramachandran plot

- B factors
  - Contributions from static and dynamic disorder
    - Well determined ~10-20 Å^2, intermediate ~20-30 Å^2, flexible 30-50 Å^2, invisible >60 Å^2.



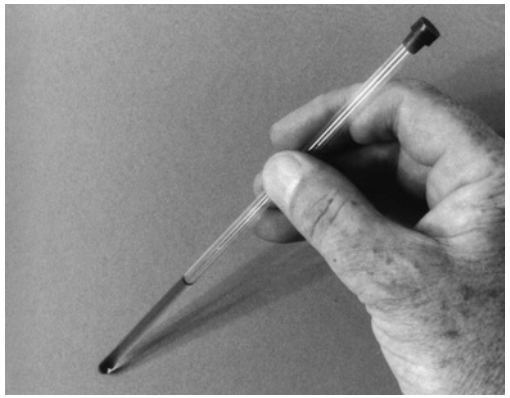
NMR Basics

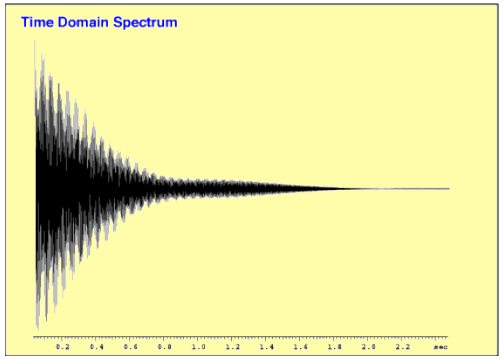
- NMR is
  - nuclear magnetic resonance
  - done on proteins IN SOLUTION
  - especially suited for studies of protein dynamics and folding
  - slow!

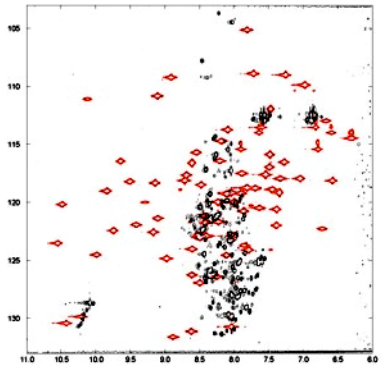
- Only certain atoms can be detected: ^{1}H, ^{13}C, ^{15}N

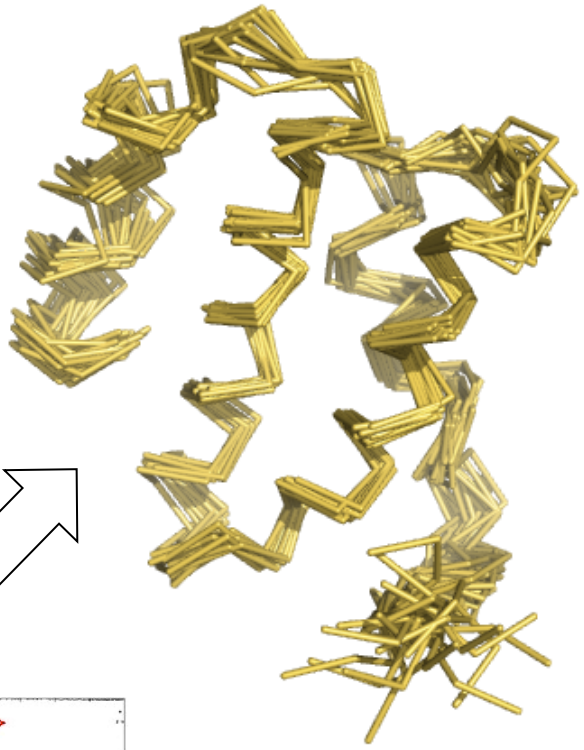
- Proteins must be
  - below 50 kDa
  - stable at high concentration (0.5-1mM) @ room temperature

NMR Spectroscopy









Evalutation of NMR Structures

- Atomic backbone RMSD: $$RMSD = \sqrt{\frac{\sum_1^n \left(x_i - \left\langle x_i' \right\rangle\right)^2}{n}}$$

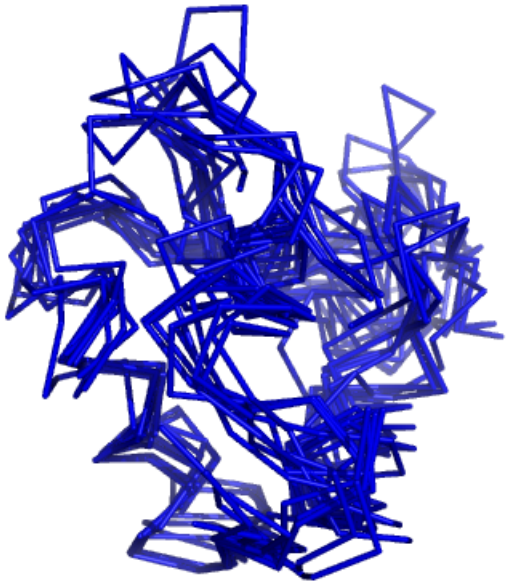
Well-defined structures

RMSDs < 0.6 Å

1T1H, Andersen et al. JBC, 2004

Less well-defined structures

RMSDs > 0.6 Å

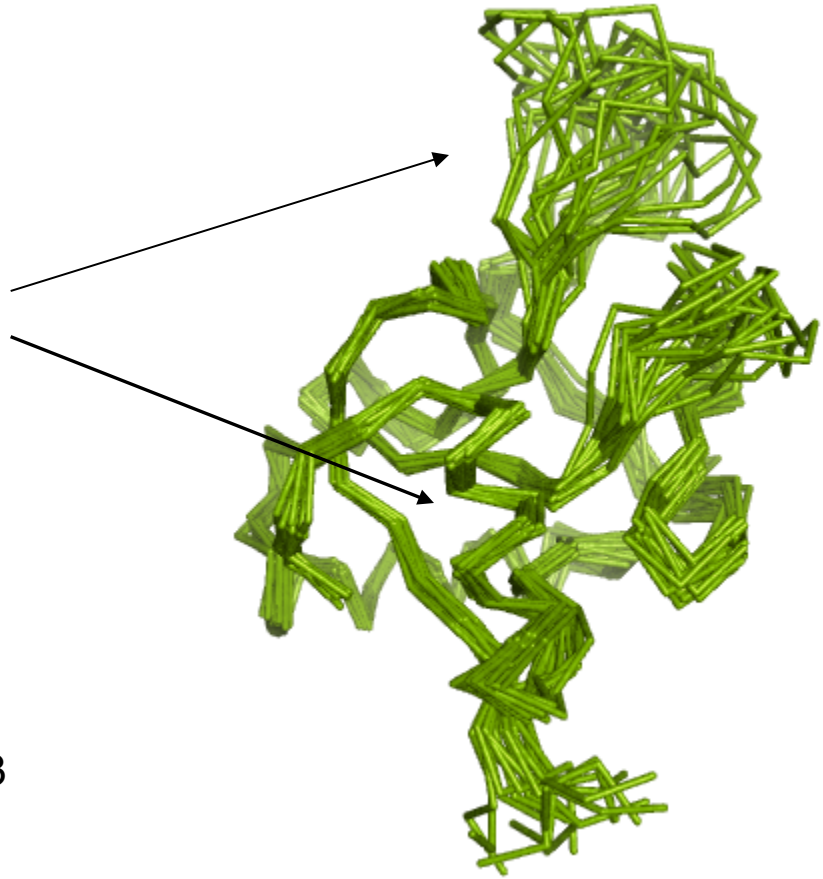


3GF1, Cooke et al. Biochemistry, 1991

Evaluation of NMR Structures

What regions in the structure are most well-defined?

Look at the pdb ensembles to see which regions are well-defined



1RJH

Nielbo et al, Biochemistry, 2003

Summary I – Protein Structure

- Proteins consist of amino acids.
- Polypeptide chains fold into specific 3D structures.
- Function is performed by the folded protein.
- Proteins are dynamic and only marginally stable.

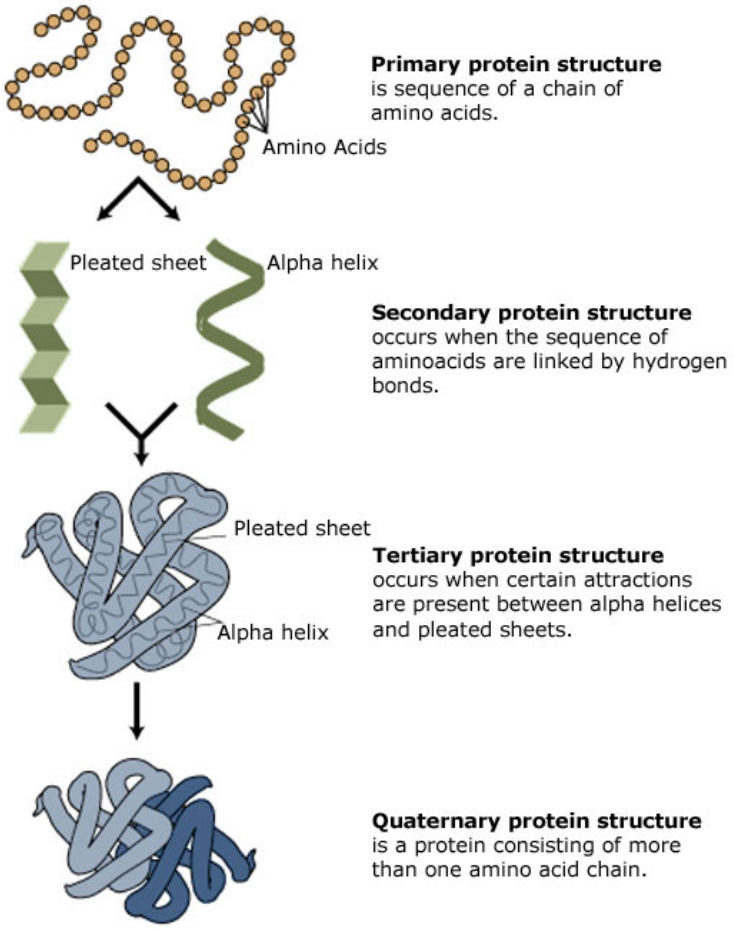


Image adapted from: National Human Genome Research Institute.

Summary

- In evolution structure is conserved longer than both function and sequence.

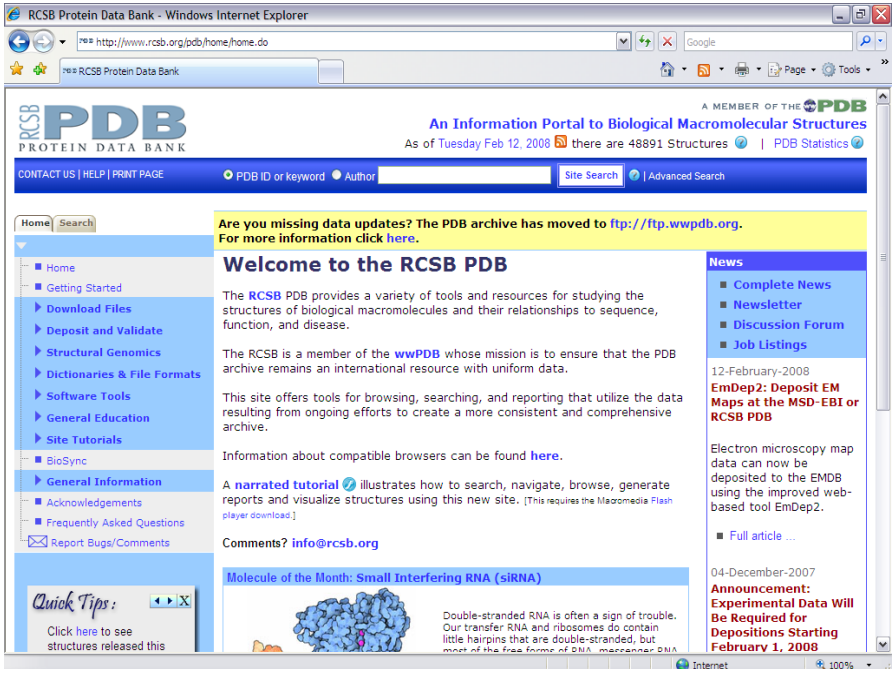
- X-ray crystallography
  - Proteins of any size
  - Proteins in crystal
  - Complete data/total map of structure
  - Many details – one model
  - Resolution, R-values, Ramachandran plot

- NMR spectroscopy
  - Proteins below 50 kDa
  - Proteins in solution
  - Incomplete data
  - Fewer details – many models
  - Restraint violations, RMSD, Ramachandran plot

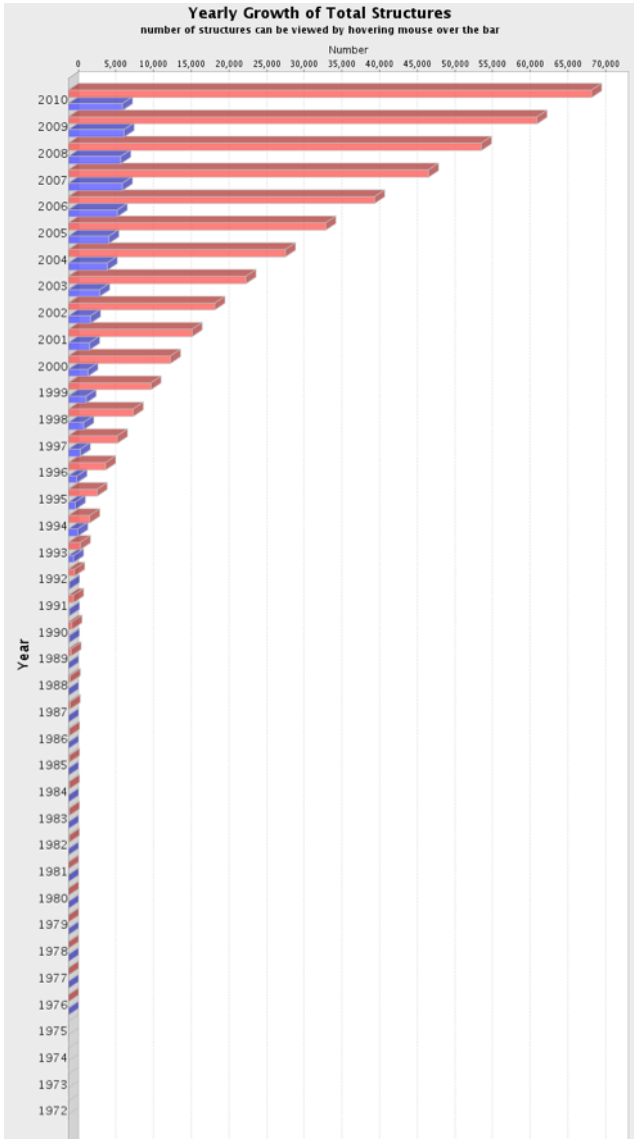
PDB
The Protein Structure Database

Protein Data Bank

- http://www.rcsb.org/
- Contents
- File structure
  - Types of structures
- Structure reports & summaries
- Quality check
- Searching
- Molecule of the Month



The Protein Data Bank



Holdings of the Protein Data Bank (PDB):

	Sep. 2001	May 2006	Nov. 2010
X-ray	13116	30860	60342
NMR	2451	5368	8666
Other	338	200	502
Total	**15905**	**36428**	**69510**

The PDB also contains nucleotide and nucleotide analogue structures.

PDB File Atom Coordinates


```
ATOM   1      N     THR A 1 25.200 26.068 37.670 1.00 25.43 N
ATOM   2      CA    THR A 1 26.443 26.547 37.135 1.00 16.70 C
ATOM   3      C     THR A 1 27.568 25.589 37.431 1.00 13.12 C
ATOM   4      O     THR A 1 27.577 25.073 38.554 1.00 15.92 O
ATOM   5      CB    THR A 1 26.745 27.891 37.843 1.00 20.41 C
ATOM   6      OG1   THR A 1 25.564 28.674 37.550 1.00 26.40 O
ATOM   7      CG2   THR A 1 27.995 28.594 37.359 1.00 22.25 C
```

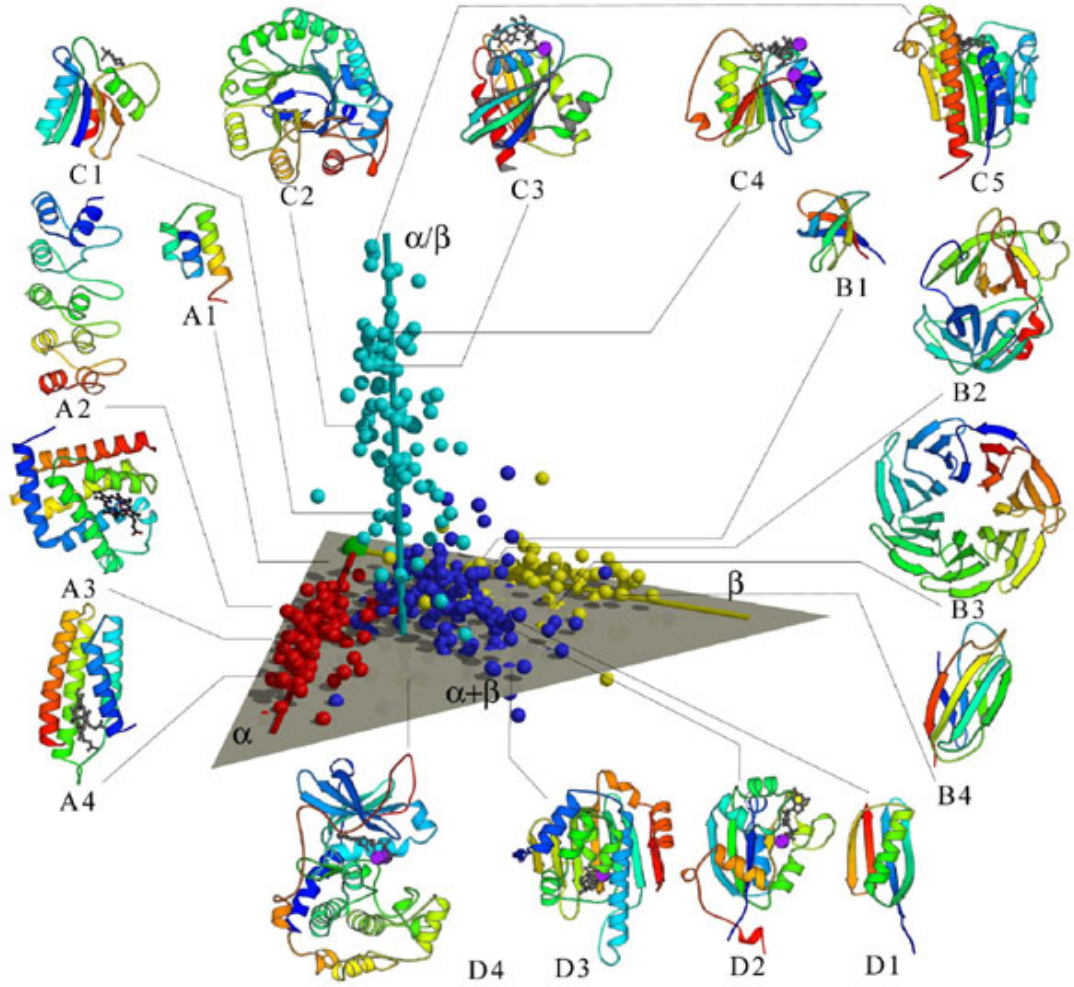
PDB File Fields

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM"	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainid	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

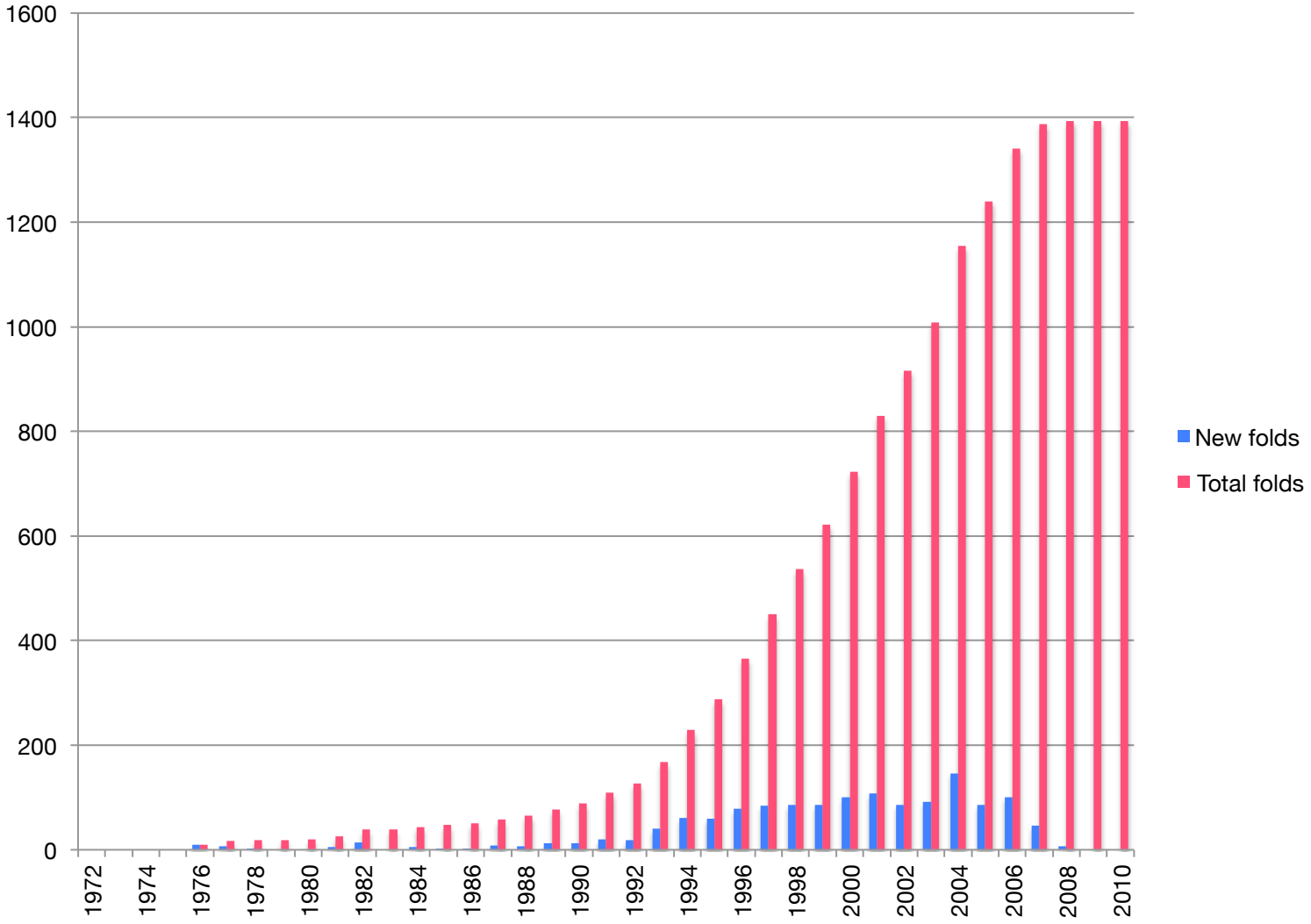
Structural Genomics

- ”Fold space coverage”
- Complete genomes
  - Disease-causing organisms
  - Model organisms
- Membrane proteins
- Protein-ligand interactions



Hou *et* al., PNAS 2003, **100**: 2386-2390

Protein Folds in PDB



Protein Structure and Visualisation

Introduction to PyMOL